Revealing Common Statistical Behaviors in Heterogeneous Populations

Andrey Zhitnikov [1] **Rotem Mulayoff** [1] **Tomer Michaeli** [1]

Abstract

In many areas of neuroscience and biological data analysis, it is desired to reveal common patterns among a group of subjects. Such analyses play important roles *e.g.,* in detecting functional brain networks from fMRI scans and in identifying brain regions which show increased activity in response to certain stimuli. Group level techniques usually assume that all subjects in the group behave according to a single statistical model, or that deviations from the common model have simple parametric forms. Therefore, complex subject-specific deviations from the common model severely impair the performance of such methods. In this paper, we propose nonparametric algorithms for estimating the common covariance matrix and the common density function of several variables in a heterogeneous group of subjects. Our estimates converge to the true model as the number of subjects tends to infinity, under very mild conditions. We illustrate the effectiveness of our methods through extensive simulations as well as on real-data from fMRI scans and from arterial blood pressure and photoplethysmogram measurements.

1. Introduction

Revealing common statistical behaviors among a group of subjects is fundamental to neuroscience and bio-medical data analysis. For example, in functional magnetic resonance imaging (fMRI) research (Bullmore et al., 1996; Smith et al., 2011; Varoquaux & Craddock, 2013), group level analyses are used for detecting brain networks from resting-state recordings (Fox et al., 2005), for detecting activities of specific regions in response to various stimuli (Haxby et al., 2001), for studying the connectivity of a specific brain region to other regions through seed based analysis (Hagler et al., 2006), etc. Group analyses often rely on the assumption that all subjects in the group behave according to the same statistical model. For example, to estimate the covariance (or partial covariance) matrix of several variables, a popular approach is to average the covariance matrices estimated for each of the individual subjects in the group (Power et al., 2011). This is done using either the Euclidean mean (arithmetic average) or the intrinsic (Riemannian) mean (Förstner & Moonen, 2003), (Fletcher & Joshi, 2007), which respects the geometry of the manifold of positive definite matrices (Varoquaux et al., 2010a).

Real data, however, rarely conform to this assumption. Often times, each subject in a group deviates from the common model *in a different way*. For example, it has been shown that estimates of connectivity patterns from fMRI scans, tend to vary significantly between subjects (Moussa et al., 2012). Subject-specific deviations may even be more dominant than the common model itself. Therefore, if ignored, these deviations may severely degrade the quality of the estimate of the common model. This phenomenon is illustrated in Fig. 1 in the context of nonparametric density estimation of two variables (brain regions). In this example, the deviations from the common model are additive and have a different distribution for each subject. Thus, as can be seen on the right, kernel density estimation (KDE) applied to the entire group, fails to reveal the common behavior.

Approaches for accounting for subject-specific deviations often make limiting assumptions. For example, in the context of covariance estimation, (Varoquaux et al., 2010b) assumed that the precision matrices of all subjects in the group have the same sparsity pattern, and proposed a modified graph Lasso technique (Friedman et al., 2008) for simultaneously estimating those matrices. In (Marrelec et al., 2006), the authors assumed that each subject's samples follow a Gaussian distribution with a covariance matrix that follows an inverse Wishart distribution around the group covariance. In the context of regression, a popular strategy is to use a linear mixed-effects model (Friston et al., 2005; Chen et al., 2013), which relies on a Gaussian distribution assumption for the subject specific factors. Similar lines of work include group-level independent component analysis (ICA) (Calhoun et al., 2001; Beckmann & Smith, 2005; Varoquaux et al., 2010c), dictionary learning (Varoquaux et al., 2011; Mensch et al., 2016), and causal structure estimation (Ramsey et al., 2010).

[1]Electrical Engineering Dept., Technion, Israel. Correspondence to: Andrey Zhitnikov <andreyz@campus.technion.ac.il>, Rotem Mulayoff <smulayof@campus.technion.ac.il>, Tomer Michaeli <tomer.m@ee.technion.ac.il>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

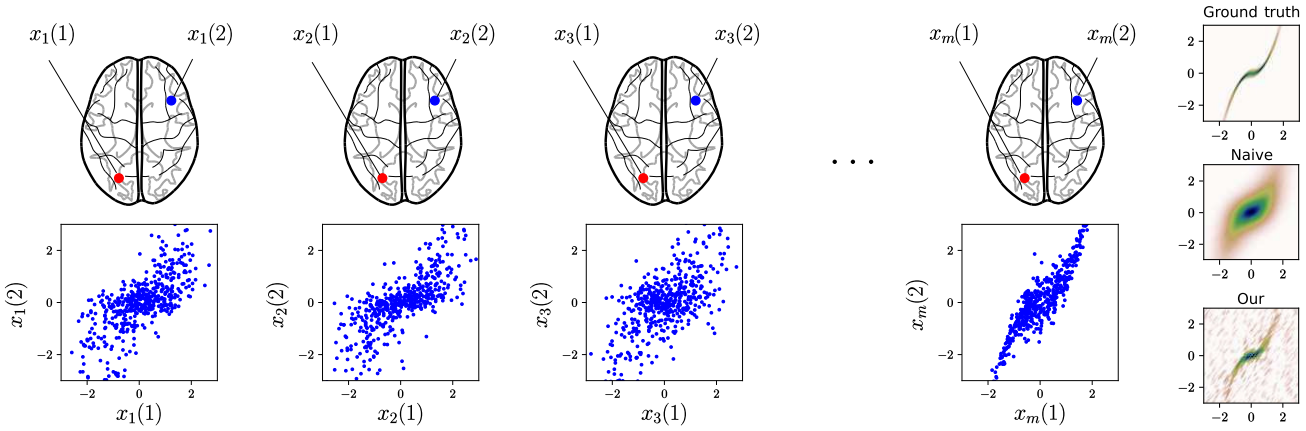

Figure 1. **The common model estimation problem.** Multiple measurements are collected from a group of subjects (500 in this simulation). The samples of the jth subject are viewed as realizations of a random vector $\boldsymbol{x}_j$, assumed to decompose as $\boldsymbol{u} + \boldsymbol{v}_j$. The random vector $\boldsymbol{u}$ is a component common to all subjects (pdf shown at the top right), whereas $\{\boldsymbol{v}_j\}$ are subject-specific factors, each of which may have a different distribution. The goal is to estimate statistical properties of the common component $\boldsymbol{u}$. As opposed to naive density estimation based on all samples from all subjects (middle right), our approach manages to unveil the fine structure of the common pdf (bottom right).

In this paper we present *non-parametric* methods for estimating a common model in the presence of subject-specific noise factors. Specifically, we present a common-covariance estimation algorithm and a common probability density function (pdf) estimation method, both of which do not assume any particular form for the underlying distributions. Our only assumption is that the subject-specific noise factors are additive and have diverse distributions (otherwise they could be considered part of the common model). In this setting, the Euclidean and Riemannian mean estimates do not approach the true covariance matrix as the number of subjects grows. In contrast, we prove that our estimate does converge to the true covariance under very mild assumptions. We verify the advantages of our approach through extensive experiments on simulated and on real data.

2. Problem formulation

Let $\boldsymbol{u} \in \mathbb{R}^d$ be a random vector, which represents the common source of variability across a group of subjects. For example, in Fig. 1, $\boldsymbol{u} \in \mathbb{R}^2$ is distributed according to the 'ground truth' density function (top right). Let $\boldsymbol{x}_j \in \mathbb{R}^d$ be a random vector, which represents the jth subject in the group (in Fig. 1, the jth scatter plot shows realizations of $\boldsymbol{x}_j$). We assume the additive model

$$\boldsymbol{x}_j = \boldsymbol{u} + \boldsymbol{v}_j, \tag{1}$$

where $\{\boldsymbol{v}_j\}$ are random vectors that are independent of $\boldsymbol{u}$ and represent subject-specific factors. Generally, each $\boldsymbol{v}_j$ has a different distribution (had they been distributed identically, they would have been part of the common model).

Given realizations of $\boldsymbol{x}_j$, for $j = 1 \dots m$, our goal is to estimate statistical properties of the common component $\boldsymbol{u}$. In particular, we are interested in estimating either the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{u}}$ or the full pdf $p_{\boldsymbol{u}}$ of $\boldsymbol{u}$.

Obviously, the performance in those estimation tasks will generally depend on both the number of subjects m and the number of samples per subject. However, here, we are interested in the common situation in which the number of samples per subject suffices to obtain reasonably accurate estimates for $\boldsymbol{\Sigma}_{\boldsymbol{x}_j}$ or $p_{\boldsymbol{x}_j}$ (*e.g.,* when the dimension d is relatively small). Our assumption is thus that the covariances (or pdfs) of the subjects $\boldsymbol{x}_j$ are known and our focus is on the problem of recovering the common covariance (or pdf) from them. To apply our algorithms in practice, one has to plug in estimates of the covariances (or pdfs) of the subjects (obtained using, *e.g.,* KDE).

3. Common covariance estimation

Since $\boldsymbol{u}$ and $\boldsymbol{v}_j$ are independent, we have from (1) that

$$\boldsymbol{\Sigma}_{\boldsymbol{x}_j} = \boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}_j} \tag{2}$$

for every $j = 1, \dots, m$. We would like to estimate the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{u}}$ of the common component, given the covariance matrices $\{\boldsymbol{\Sigma}_{\boldsymbol{x}_j}\}$ of the subjects. To avoid ambiguity, we define the common component $\boldsymbol{\Sigma}_{\boldsymbol{u}}$ to be the largest one satisfying such a decomposition. In particular, this means that the smallest eigenvalue of (at least some of) the subject-specific factors $\{\boldsymbol{\Sigma}_{\boldsymbol{v}_j}\}$ must be arbitrarily small. Indeed, otherwise there would exist some $\alpha > 0$ such that $\boldsymbol{\Sigma}_{\boldsymbol{v}_j} \succ \alpha \boldsymbol{I}$ for every j so that $\alpha \boldsymbol{I}$ would be common to all $\{\boldsymbol{\Sigma}_{\boldsymbol{v}_j}\}$ and not subject-specific. In other words, the common component in this case is in fact $\boldsymbol{\Sigma}_{\boldsymbol{u}} + \alpha \boldsymbol{I}$ and the noise covariances are $\boldsymbol{\Sigma}_{\boldsymbol{v}_j} - \alpha \boldsymbol{I}$.

Let us first informally describe the key idea underlying

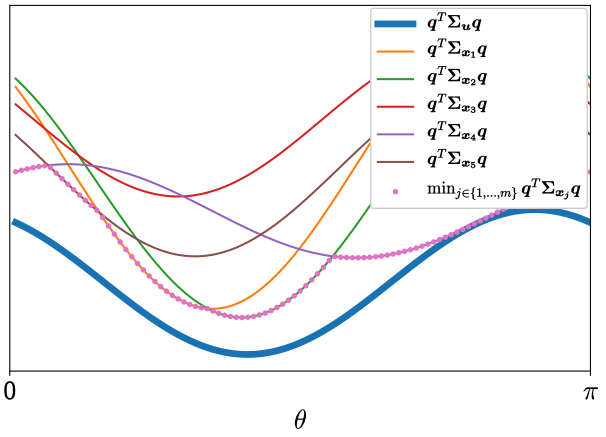

Figure 2. **Visualization of the optimization problems (5), (6) for 2×2 matrices.** The quadratic forms corresponding to the common covariance (thick line), the covariances of the subjects (thin lines) and their pointwise minimum (dots), are shown as functions of the angle θ of the 2D vector $\boldsymbol{q} = (\cos(\theta)\ \ \sin(\theta))^T$.

our method, and then provide a formal "group consistency" result. Denote the eigenvalues of $\boldsymbol{\Sigma_u}$ by $\lambda_1 \le \lambda_2 \le \ldots, \lambda_d$ and the corresponding eigenvectors by $\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_d$. We will begin by estimating the smallest eigenvalue, λ_1, and its associated eigenvector, $\boldsymbol{q}_1$. By definition,

$$\lambda_1 = \min_{\|\boldsymbol{q}\|=1} \boldsymbol{q}^T \boldsymbol{\Sigma_u} \boldsymbol{q}. \tag{3}$$

Now, observe that

$$\boldsymbol{q}^T \boldsymbol{\Sigma_u} \boldsymbol{q} \le \boldsymbol{q}^T \boldsymbol{\Sigma}_{\boldsymbol{x}_j} \boldsymbol{q}, \quad \forall j, \forall \boldsymbol{q} \tag{4}$$

since $\boldsymbol{q}^T \boldsymbol{\Sigma}_{\boldsymbol{x}_j} \boldsymbol{q} = \boldsymbol{q}^T \boldsymbol{\Sigma_u} \boldsymbol{q} + \boldsymbol{q}^T \boldsymbol{\Sigma}_{\boldsymbol{v}_j} \boldsymbol{q}$ and $\boldsymbol{q}^T \boldsymbol{\Sigma}_{\boldsymbol{v}_j} \boldsymbol{q} \ge 0$. Our assumption, which we formalize mathematically below, is that the subject-specific noise covariances $\boldsymbol{\Sigma}_{\boldsymbol{v}_j}$ are diverse in the sense that their bottom eigenvectors tend to point in different directions. This, together with the fact their smallest eigenvalue can be arbitrarily small, implies that as the number of subjects grows, it becomes increasingly likely that for every direction $\boldsymbol{q}$, at least one of the values $\{\boldsymbol{q}^T \boldsymbol{\Sigma}_{\boldsymbol{v}_j} \boldsymbol{q}\}_{j=1}^m$ be small. This motivates us to estimate λ_1 and $\boldsymbol{q}_1$ as

$$\hat{\boldsymbol{q}}_1 = \arg\min_{\|\boldsymbol{q}\|=1} \min_{j\in\{1,\ldots,m\}} \boldsymbol{q}^T \boldsymbol{\Sigma}_{\boldsymbol{x}_j} \boldsymbol{q}, \tag{5}$$

$$\hat{\lambda}_1 = \min_{\|\boldsymbol{q}\|=1} \min_{j\in\{1,\ldots,m\}} \boldsymbol{q}^T \boldsymbol{\Sigma}_{\boldsymbol{x}_j} \boldsymbol{q}. \tag{6}$$

That is, we minimize over the pointwise minimum of the quadratic functions of the individual subjects. Figure 2 visualizes this objective for the case of 2×2 matrices. Here, the thick blue curve corresponds to the desired objective function (3), which we cannot directly minimize (as it involves the unknown $\boldsymbol{\Sigma_u}$). The thin curves correspond to the quadratic functions of the subjects (involving the known matrices $\{\boldsymbol{\Sigma}_{\boldsymbol{x}_j}\}$). As can be seen, the pointwise minimum of the thin curves (dotted curve) is close to the thick curve when the number of subjects is large.

Algorithm 1 Common covariance estimation

Input: Covariance matrices $\boldsymbol{\Sigma}_{\boldsymbol{x}_1}, \ldots, \boldsymbol{\Sigma}_{\boldsymbol{x}_m}$ in $\mathbb{R}^{d\times d}$.
Output: Common covariance estimate $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}$.
for $k = 1 \ldots d$ **do**
Using (14), compute $\hat{\boldsymbol{q}}_k$ and $\hat{\lambda}_k$ as

$$\hat{\boldsymbol{q}}_k = \arg\min_{\boldsymbol{q}\in\mathcal{S}_k} \min_{j\in\{1,\ldots,m\}} \boldsymbol{q}^T \boldsymbol{\Sigma}_{\boldsymbol{x}_j} \boldsymbol{q}, \tag{11}$$

$$\hat{\lambda}_k = \min_{\boldsymbol{q}\in\mathcal{S}_k} \min_{j\in\{1,\ldots,m\}} \boldsymbol{q}^T \boldsymbol{\Sigma}_{\boldsymbol{x}_j} \boldsymbol{q}, \tag{12}$$

where

$$\mathcal{S}_k = \left\{\boldsymbol{q} : \|\boldsymbol{q}\| = 1,\ \boldsymbol{q} \perp \mathrm{span}\{\hat{\boldsymbol{q}}_1, \ldots, \hat{\boldsymbol{q}}_{k-1}\}\right\}. \tag{13}$$

end for
Construct $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}$ from $\{\hat{\boldsymbol{q}}_k\}_{k=1}^m$ and $\{\hat{\lambda}_k\}_{k=1}^m$ as in (10).

Next, we turn to estimate λ_2 and $\boldsymbol{q}_2$. Note that

$$\begin{aligned}\lambda_2 &= \min_{\{q:\|\boldsymbol{q}\|=1, \boldsymbol{q}\perp\boldsymbol{q}_1\}} \boldsymbol{q}^T \boldsymbol{\Sigma_u} \boldsymbol{q} \\ &\le \min_{\{q:\|\boldsymbol{q}\|=1, \boldsymbol{q}\perp\boldsymbol{q}_1\}} \boldsymbol{q}^T \boldsymbol{\Sigma}_{\boldsymbol{x}_j} \boldsymbol{q}, \quad \forall j. \end{aligned} \tag{7}$$

Therefore, following the logic above, and replacing $\boldsymbol{q}_1$ by its estimate $\hat{\boldsymbol{q}}_1$, we propose to calculate $\hat{\boldsymbol{q}}_2$ and $\hat{\lambda}_2$ as

$$\hat{\boldsymbol{q}}_2 = \arg\min_{\{\boldsymbol{q}:\|\boldsymbol{q}\|=1, \boldsymbol{q}\perp\hat{\boldsymbol{q}}_1\}} \min_{j\in\{1,\ldots,m\}} \boldsymbol{q}^T \boldsymbol{\Sigma}_{\boldsymbol{x}_j} \boldsymbol{q}, \tag{8}$$

$$\hat{\lambda}_2 = \min_{\{\boldsymbol{q}:\|\boldsymbol{q}\|=1, \boldsymbol{q}\perp\hat{\boldsymbol{q}}_1\}} \min_{j\in\{1,\ldots,m\}} \boldsymbol{q}^T \boldsymbol{\Sigma}_{\boldsymbol{x}_j} \boldsymbol{q}. \tag{9}$$

This process can be repeated, where at the kth step, we constrain the search to the subspace orthogonal to $\mathrm{span}\{\hat{\boldsymbol{q}}_1, \ldots, \hat{\boldsymbol{q}}_{k-1}\}$. The last eigenvector, $\hat{\boldsymbol{q}}_d$, is completely determined by $\hat{\boldsymbol{q}}_1, \ldots, \hat{\boldsymbol{q}}_{d-1}$ and thus does not involve an optimization problem. The associated eigenvalue is estimated as $\hat{\lambda}_d = \min_{j\in\{1,\ldots,m\}} \hat{\boldsymbol{q}}_d^T \boldsymbol{\Sigma}_{\boldsymbol{x}_j} \hat{\boldsymbol{q}}_d$.

Having estimated all the eigenvalues and eigenvectors, we construct our estimate of $\boldsymbol{\Sigma_u}$ as

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}} = \sum_{k=1}^{d} \hat{\lambda}_k \hat{\boldsymbol{q}}_k \hat{\boldsymbol{q}}_k^T. \tag{10}$$

This is summarized in Alg. 1.

3.1. Practical implementation

The objective in Problems (11),(12) is the pointwise minimum of a finite set of continuous (quadratic) functions over

a compact set. Therefore, the minimum is attained at the minimum of one of those functions, each of which has a closed form. Specifically, when $k = 1$, we only have the constraint $\|\boldsymbol{q}\| = 1$, and the minimum of the jth problem is the smallest eigenvalue of $\boldsymbol{\Sigma}_{\boldsymbol{x}_j}$ (attained by the corresponding eigenvector). When $k > 1$, we have an additional set of linear constraints, which can be written as $\boldsymbol{Q}_k \boldsymbol{q} = 0$, where $\boldsymbol{Q}_k = \sum_{i=1}^{k-1} \hat{\boldsymbol{q}}_i \hat{\boldsymbol{q}}_i^T$. In this case, the minimizer is given by the top eigenvector of $(\boldsymbol{I} - \boldsymbol{Q}_k)(c\boldsymbol{I} - \boldsymbol{\Sigma}_{\boldsymbol{x}_j})(\boldsymbol{I} - \boldsymbol{Q}_k)$, which we denote by $\boldsymbol{v}_j^k$, where c is any constant such that $c\boldsymbol{I} - \boldsymbol{\Sigma}_{\boldsymbol{x}_j} \succ 0$ (Blau & Michaeli, 2017). Thus, in summary,

$$\hat{\boldsymbol{q}}_k = \boldsymbol{v}_{j^*}^k, \qquad \hat{\lambda}_k = (\boldsymbol{v}_{j^*}^k)^T \boldsymbol{\Sigma}_{\boldsymbol{x}_j} \boldsymbol{v}_{j^*}^k, \tag{14}$$

where $j^* = \arg\min_{j \in \{1,\dots,m\}} (\boldsymbol{v}_j^k)^T \boldsymbol{\Sigma}_{\boldsymbol{x}_j} \boldsymbol{v}_j^k$.

In the Supplementary Material, we discuss ways to speed up the estimation on parallel platforms.

3.2. Group consistency

To analyze the behavior of Alg. 1 as the number of subjects m increases, one must assume something regarding the variability of the subject-specific noise covariances $\boldsymbol{\Sigma}_{\boldsymbol{v}_j}$. A rather general assumption is that they are independent draws from some distribution over PSD matrices, namely

$$\boldsymbol{\Sigma}_{\boldsymbol{v}_j} \sim p_{\boldsymbol{\Sigma}_{\boldsymbol{v}}}. \tag{15}$$

The next theorem shows that under very mild conditions on $p_{\boldsymbol{\Sigma}_{\boldsymbol{v}}}$, our estimate $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}$ converges to $\boldsymbol{\Sigma}_{\boldsymbol{u}}$ almost surely (a.s.). We refer to this as *group consistency*.

Theorem 1 (Group consistency)**.** *Assume that*

$$\mathbb{P}\left(\lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{v}}) \le \alpha\right) = 1 \tag{16}$$

for some $\alpha > 0$ and that

$$\mathbb{P}\left(\boldsymbol{q}^T \boldsymbol{\Sigma}_{\boldsymbol{v}} \boldsymbol{q} \le \epsilon\right) > 0 \tag{17}$$

for every $\epsilon > 0$ and every unit-norm $\boldsymbol{q}$. Let $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^m$ denote the estimate produced by Alg. 1 using m subjects. Then

$$\mathbb{P}\left(\lim_{m\to\infty} \left\|\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^m - \boldsymbol{\Sigma}_{\boldsymbol{u}}\right\| = 0\right) = 1. \tag{18}$$

Assumption (16) merely states that the noise factors are not arbitrarily large. Assumption (17) is a condition on the distribution of the smallest eigenvalue of $\boldsymbol{\Sigma}_{\boldsymbol{v}}$ and its associated eigenvector. Roughly speaking, it requires that there be a positive probability for the smallest eigenvalue to be arbitrarily small and, simultaneously, for the corresponding eigenvector to point in any direction (*i.e.,* this eigenvector can have any distribution on the unit sphere as long as it does not vanish on a set of nonzero Lebesgue measure). Recall that the condition on the smallest eigenvalue is actually part of the definition of the common covariance estimation problem, and therefore not a limiting assumption.

To prove the theorem, let us denote $\psi(\boldsymbol{q}) \triangleq \boldsymbol{q}^T \boldsymbol{\Sigma}_{\boldsymbol{u}} \boldsymbol{q}$, $g_j(\boldsymbol{q}) \triangleq \boldsymbol{q}^T \boldsymbol{\Sigma}_{\boldsymbol{v}_j} \boldsymbol{q}$, and $h_m(\boldsymbol{q}) \triangleq \min_{j\in\{1,\dots,m\}} g_j(\boldsymbol{q})$. Note that $\psi(\boldsymbol{q})$ is a deterministic function (as $\boldsymbol{\Sigma}_{\boldsymbol{u}}$ is deterministic) whereas $\{g_j(\boldsymbol{q})\}$ and $\{h_m(\boldsymbol{q})\}$ are sequences of random functions (as $\{\boldsymbol{\Sigma}_{\boldsymbol{v}_j}\}$ are random). We will need the following lemmas (see proofs in the Supplementary).

Lemma 1. *For every $\boldsymbol{q}$, the sequence of random variables $\{h_m(\boldsymbol{q})\}$ converges to zero almost surely. Furthermore, for any sequence of vectors $\{\boldsymbol{q}_m\}_{m=1}^{\infty}$ that converges to some vector $\boldsymbol{q}^*$, the sequence of random variables $\{h_m(\boldsymbol{q}_m)\}$ converges to zero almost surely.*

Lemma 2. *Let $\phi(\boldsymbol{q})$ be a continuous bounded function on a compact set $\mathcal{C}$, which achieves a strict global minimum at $\boldsymbol{q}^* \in \mathcal{C}$. Let $\{f_n(\boldsymbol{q})\}_{n=1}^{\infty}$ be a sequence of continuous bounded nonnegative functions on $\mathcal{C}$ satisfying $f_n(\boldsymbol{q}^*) \to 0$, and let $w_n(\boldsymbol{q}) = \phi(\boldsymbol{q}) + f_n(\boldsymbol{q})$. Then any sequence of the form $\boldsymbol{q}_n \in \arg\min_{\boldsymbol{q}\in\mathcal{C}} w_n(\boldsymbol{q})$ converges to $\boldsymbol{q}^*$, and the sequence $w_n(\boldsymbol{q}_n)$ converges to $\phi(\boldsymbol{q}^*)$.*

proof of Theorem 1. For simplicity, we prove the theorem for $d = 2$. The extension to higher dimensions is similar.

Since problem (11) is symmetric, we can divide the unit circle into two disjoint half circles $\mathcal{A}$ and $\mathcal{B}$ such that $\mathcal{A}$ is closed, and restrict the search for the minimum to $\mathcal{A}$. Let us first assume that $\lambda_1 \neq \lambda_2$. In this case, the minimum of $\psi(\boldsymbol{q})$ over the unit circle is achieved at the points $\boldsymbol{q}_1$ and $-\boldsymbol{q}_1$. Without loss of generality, we assume that $\boldsymbol{q}_1 \in \mathcal{A}$ and $-\boldsymbol{q}_1 \in \mathcal{B}$. The objective in (11) can be written as $\psi(\boldsymbol{q}) + h_m(\boldsymbol{q})$. Since $h_m(\boldsymbol{q})$ is continuous for every m and $h_m(\boldsymbol{q}_1) \overset{\text{a.s.}}{\to} 0$ (Lemma 1), the conditions of Lemma 2 hold a.s. Therefore, our estimate of the bottom eigenvector, $\hat{\boldsymbol{q}}_1^m$, converges a.s. to the true eigenvector $\boldsymbol{q}_1$, namely

$$\hat{\boldsymbol{q}}_1^m \overset{\text{a.s.}}{\to} \boldsymbol{q}_1. \tag{19}$$

Our estimate (12) of the bottom eigenvalue, $\hat{\lambda}_1^m$, can be written as $\psi(\hat{\boldsymbol{q}}_1^m) + h_m(\hat{\boldsymbol{q}}_1^m)$. Since $\hat{\boldsymbol{q}}_1^m \overset{\text{a.s.}}{\to} \boldsymbol{q}_1$, we have from Lemma 1 that $h_m(\hat{\boldsymbol{q}}_1^m) \overset{\text{a.s.}}{\to} 0$, and therefore $\hat{\lambda}_1^m \overset{\text{a.s.}}{\to} \psi(\boldsymbol{q}_1) = \lambda_1$ as well.

The top eigenvector is given by $\boldsymbol{q}_2 = \boldsymbol{R}\boldsymbol{q}_1$, where $\boldsymbol{R}$ is a $90°$ rotation matrix, and our estimate of this eigenvector is simply $\hat{\boldsymbol{q}}_2 = \boldsymbol{R}\hat{\boldsymbol{q}}_1$. Therefore, (19) implies that also

$$\hat{\boldsymbol{q}}_2^m \overset{\text{a.s.}}{\to} \boldsymbol{q}_2. \tag{20}$$

The convergence of $\hat{\lambda}_2^m$ to λ_2 follows similarly by Lemma 1.

Let us now treat the case where $\lambda_1 = \lambda_2$. In this setting, the vectors $\hat{\boldsymbol{q}}_1^m, \hat{\boldsymbol{q}}_2^m$ do not necessarily converge. However, for the matrix $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^m$ to converge to $\boldsymbol{\Sigma}_{\boldsymbol{u}}$, it suffices that only

the eigenvalue estimates $\hat{\lambda}_1^m, \hat{\lambda}_2^m$ converge to λ_1, λ_2 (in that case, the vectors $\hat{\boldsymbol{q}}_1^m, \hat{\boldsymbol{q}}_2^m$ have no effect in (10)). To see that the eigenvalues converge, note that the solution of (12) is bounded from below by $\min_{\boldsymbol{q}\in\mathcal{S}_1}\psi(\boldsymbol{q}) = \lambda_1$, because $h_m(\boldsymbol{q}) \geq 0$. Additionally, we have that

$$\begin{aligned}\hat{\lambda}_1^m &= \min_{\boldsymbol{q}\in\mathcal{S}_1}\min_{j\in\{1,\ldots,m\}} \boldsymbol{q}^T\boldsymbol{\Sigma}_{\boldsymbol{x}_j}\boldsymbol{q} \\ &= \lambda_1 + \min_{\boldsymbol{q}\in\mathcal{S}_1} h_m(\boldsymbol{q}) \\ &\leq \lambda_1 + h_m(\bar{\boldsymbol{q}}) \overset{\text{a.s.}}{\to} \lambda_1, \end{aligned} \tag{21}$$

where $\bar{\boldsymbol{q}}$ is an arbitrary point in $\mathcal{S}_1$, and the convergence is due to Lemma 1. Therefore $\hat{\lambda}_1^m$ converges to λ_1. Similar arguments can be invoked to show that $\hat{\lambda}_2^m$ converges to λ_2.

Since the eigenvectors and eigenvalues converge, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^m$ converges to $\boldsymbol{\Sigma}_{\boldsymbol{u}}$, and the proof is complete. ■

4. Common density function estimation

Next, we address the problem of estimating the pdf $p_{\boldsymbol{u}}$ of the common component, given the pdfs $\{p_{\boldsymbol{x}_j}\}$ of the subjects in the group.

Since $\boldsymbol{u}$ and $\boldsymbol{x}_j$ are statistically independent, we have that

$$p_{\boldsymbol{x}_j}(\boldsymbol{\alpha}) = \left(p_{\boldsymbol{u}} * p_{\boldsymbol{v}_j}\right)(\boldsymbol{\alpha}), \tag{22}$$

where '$*$' denotes convolution. Furthermore,

$$\varphi_{\boldsymbol{x}_j}(\boldsymbol{t}) = \varphi_{\boldsymbol{u}}(\boldsymbol{t})\,\varphi_{\boldsymbol{v}_j}(\boldsymbol{t}), \tag{23}$$

where $\varphi_{\boldsymbol{z}}(\boldsymbol{t}) = \mathbb{E}[e^{j\boldsymbol{t}^T\boldsymbol{z}}]$ denotes the characteristic function of a random vector $\boldsymbol{z}$. We will focus on estimating $\varphi_{\boldsymbol{u}}(\boldsymbol{t})$, from which $p_{\boldsymbol{u}}$ can be retrieved by a Fourier transform.

A well known property of characteristic functions is that $|\varphi_{\boldsymbol{z}}(\boldsymbol{t})| \leq 1$ for every $\boldsymbol{t}$. Therefore, we have from (23) that $|\varphi_{\boldsymbol{u}}(\boldsymbol{t})| \geq |\varphi_{\boldsymbol{x}_j}(\boldsymbol{t})|$ for every j and for all $\boldsymbol{t}$. In particular,

$$|\varphi_{\boldsymbol{u}}(\boldsymbol{t})| \geq \max_{j\in\{1,\ldots,m\}} \left|\varphi_{\boldsymbol{x}_j}(\boldsymbol{t})\right|, \quad \forall \boldsymbol{t}. \tag{24}$$

Based on this observation, we propose to take the maximum among the values $\{|\varphi_{\boldsymbol{x}_j}(\boldsymbol{t})|\}_{j=1}^m$ as our estimate of $|\varphi_{\boldsymbol{u}}(\boldsymbol{t})|$, for every $\boldsymbol{t}$. The idea is that if the noise characteristic functions $\{\varphi_{\boldsymbol{v}_j}(\boldsymbol{t})\}$ are diverse, then for every $\boldsymbol{t}$, it is likely that at least one of them attain a value close to 1 (in absolute value). Namely, at least one of the values $\{|\varphi_{\boldsymbol{x}_j}(\boldsymbol{t})|\}$ is close to $|\varphi_{\boldsymbol{u}}(\boldsymbol{t})|$, which justifies our estimator. To estimate the phase of $\varphi_{\boldsymbol{u}}(\boldsymbol{t})$, we take the phase of the characteristic function $\varphi_{\boldsymbol{x}_j}(\boldsymbol{t})$ that attains the maximum. That is, we construct our estimate as

$$\begin{aligned} k(\boldsymbol{t}) &= \arg\max_{j\in\{1,\ldots,m\}} |\varphi_{\boldsymbol{x}_j}(\boldsymbol{t})|, \\ \hat{\varphi}_{\boldsymbol{u}}(\boldsymbol{t}) &= \varphi_{\boldsymbol{x}_{k(\boldsymbol{t})}}(\boldsymbol{t}). \end{aligned} \tag{25}$$

Algorithm 2 Common density estimation

Input: Density functions $p_{\boldsymbol{x}_1}, \ldots, p_{\boldsymbol{x}_m}$.
Output: Common density estimate $\hat{p}_{\boldsymbol{u}}$.
for $j = 1 \ldots m$ **do**
  Set $\varphi_{\boldsymbol{x}_j} \leftarrow \text{IDFT}\{p_{\boldsymbol{x}_j}\}$.
  for all $\boldsymbol{t}$ **do**
    Set k as the index of the largest value in $\{\varphi_{\boldsymbol{x}_j}(\boldsymbol{t})\}$.
    Set $\hat{\varphi}_{\boldsymbol{u}}(\boldsymbol{t}) \leftarrow \varphi_{\boldsymbol{x}_k}(\boldsymbol{t})$.
  end for
end for
Set $\hat{p}_{\boldsymbol{u}} \leftarrow \text{DFT}\{\hat{\varphi}_{\boldsymbol{u}}\}$.
Truncate the negative values of $\hat{p}_{\boldsymbol{u}}$ and normalize it to have unit area.

Note that our phase estimate is accurate when the pdfs $\{p_{\boldsymbol{v}_j}\}$ are symmetric (*e.g.*, when $\{\boldsymbol{v}_j\}$ are zero-mean Gaussian random vectors). Indeed, in that case the phase of $\varphi_{\boldsymbol{v}_j}$ is zero, so that the phase of $\varphi_{\boldsymbol{u}}$ equals the phase of $\varphi_{\boldsymbol{x}_j}$. Our common pdf estimation algorithm is summarized in Alg. 2.

It is interesting to note that Alg. 2 has been proposed in the Image Processing community, as a way of removing blur from several blurry images the of same scene (Delbracio & Sapiro, 2015). The analogy to our setting is quite natural. The functions $p_{\boldsymbol{x}_j}$ in our context can be thought of as "blurry" versions of the function $p_{\boldsymbol{u}}$, where the "blur kernels" are the functions $p_{\boldsymbol{v}_j}$ (see (22)).

5. Experiments

In this section we verify the effectiveness of our methods, first on simulated data and then on real data.

5.1. Estimation of Pearson correlation coefficient

In our first experiment, we study the behavior of our common covariance estimator as a function of the number of subjects and the signal to noise ratio (SNR). We take the common component $\boldsymbol{u}$ to be a two-dimensional random vector with covariance matrix

$$\boldsymbol{\Sigma}_{\boldsymbol{u}} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}. \tag{26}$$

Our goal is to estimate the Pearson correlation coefficient between $\boldsymbol{u}(1)$ and $\boldsymbol{u}(2)$ (which is $\rho = 0.5$ in this case) from the perturbed versions $\boldsymbol{\Sigma}_{\boldsymbol{x}_j} = \boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}_j}$. This can be done by first estimating $\boldsymbol{\Sigma}_{\boldsymbol{u}}$ and then normalizing the off-diagonal entry by the square-roots of the diagonal entries. We compare our estimator (Alg. 1) with naive averaging of $\{\boldsymbol{\Sigma}_{\boldsymbol{x}_j}\}$ using either Euclidean or Riemannian mean.

We generate the matrices $\{\boldsymbol{\Sigma}_{\boldsymbol{v}_j}\}$ as

$$\boldsymbol{\Sigma}_{\boldsymbol{v}_j} = \boldsymbol{M}_j\boldsymbol{\Lambda}_j\boldsymbol{M}_j^T, \tag{27}$$

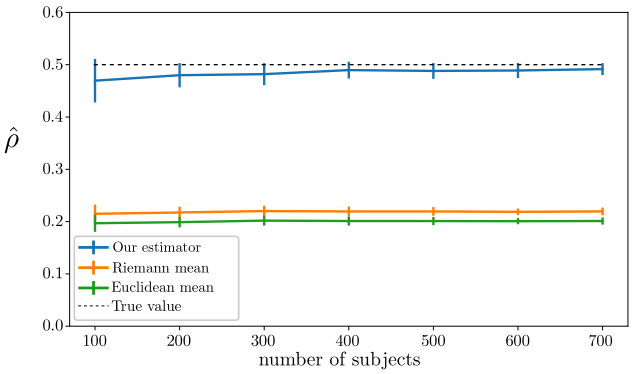

Figure 3. **Correlation estimate vs. number of subjects.** A correlation coefficient of $\rho = 0.5$ is estimated from a varying number of noisy 2×2 covariance matrices at an SNR of 0.66. Our algorithm produces accurate estimates, whereas the Euclidean and Riemannian means have severe biases.

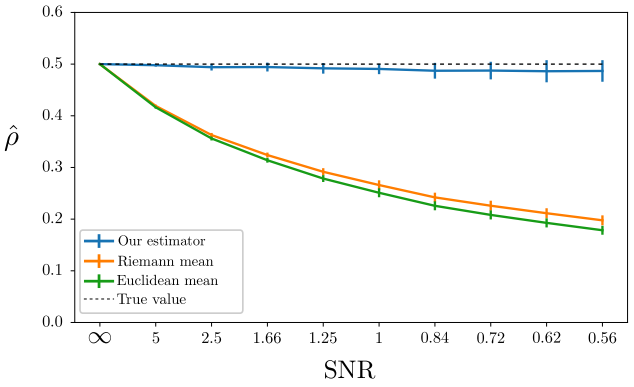

Figure 4. **Correlation estimate vs. SNR.** A correlation coefficient of $\rho = 0.5$ is estimated from 300 noisy 2×2 covariance matrices. Our algorithm produces accurate estimates even at low SNRs.

where $\boldsymbol{M}_j$ are random rotation matrices whose angles are distributed uniformly in $[0, 2\pi]$, and $\boldsymbol{\Lambda}_j$ are random diagonal matrices $\boldsymbol{\Lambda}_j = \mathrm{diag}\{\beta_1^j, \beta_2^j\}$ with $\beta_1^j \sim U[0, b]$ and $\beta_2^j \sim U[b, 2b]$ for some $b > 0$. We draw $\{\boldsymbol{M}_j\}, \{\beta_1^j\}, \{\beta_2^j\}$ independently. The SNR, which we define as $\mathrm{SNR} = \mathrm{Tr}\{\boldsymbol{\Sigma_u}\}/\mathbb{E}\{\mathrm{Tr}\{\boldsymbol{\Sigma_v}\}\}$, is $1/b$ in this case.

Figures 3 and 4 visualize the mean and variance of our estimator as well as of the naive Euclidean and Riemanian mean estimators (using 200 trials per setting) as functions of the number of subjects and the SNR. As can be seen, while the variance of our estimator is slightly larger than the variances of the naive estimators, its bias is significantly smaller. Therefore, overall, it attains a substantially lower mean square error. Figure 3 also indicates that our estimator is asymptotically (in the number of subjects) unbiased. The naive estimators, on the other hand, have severe biases, which do not decrease with the number of subjects. Figure 4 further illustrates that the performance of the naive estimators degrades rapidly as the SNR decreases, while our estimator remains relatively accurate even at low SNRs.

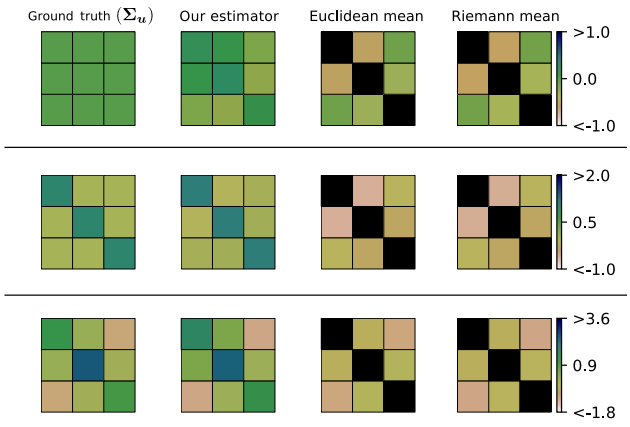

Figure 5. **Common covariance estimation with clustered noise covariances**. In each row, a different 3×3 covariance matrix (left) was estimated from 1000 noisy versions, using our method and using the Euclidean and Riemannian means. The noise covariances have a preference towards certain patterns (their eigenvectors are not distributed uniformly on the unit sphere). This causes severe bias in the naive methods, yet does not impact our algorithm.

In this example, the poor performance of the naive estimators is mainly rooted in their over-estimation of the diagonal entries of $\boldsymbol{\Sigma_u}$. This happens because the contributions of the noise matrices $\{\boldsymbol{\Sigma}_{\boldsymbol{v}_j}\}$ are only positive on the diagonal, so that averaging does not cancel them out.

5.2. Clustered subject-specific noise covariances

In most practical cases, the advantage of our approach is not confined to the diagonal elements of $\boldsymbol{\Sigma_u}$. Specifically, although our algorithm relies on the diversity of the noise covariances, it does not require their eigenvectors to be uniformly distributed on the unit sphere. Therefore, our technique can even handle cases in which the noise covariances tend to cluster around a certain matrix. As long as there exists a nonzero probability to encounter matrices away from the cluster, our algorithm is guaranteed to produce an accurate estimate as the number of subjects grows. This is in contrast to naive averaging, which typically produces estimates with severe bias in all matrix entries.

To illustrate this, we next perform a 3×3 common covariance estimation experiment. We generate $\boldsymbol{\Sigma}_{\boldsymbol{v}_j}$ as in (27), where now we construct the unitary matrix $\boldsymbol{M}_j$ as

$$\begin{pmatrix} \sin(\theta_1)\cos(\theta_2) & \sin(\theta_1)\sin(\theta_2) & \cos(\theta_1) \\ \cos(\theta_1)\cos(\theta_2) & \cos(\theta_1)\sin(\theta_2) & -\sin(\theta_1) \\ -\sin(\theta_2) & \cos(\theta_2) & 0 \end{pmatrix}, \quad (28)$$

with θ_1 and θ_2 being two independent random variables with a normal distribution $\mathcal{N}(1, 1)$ truncated to $[0, \pi]$ and

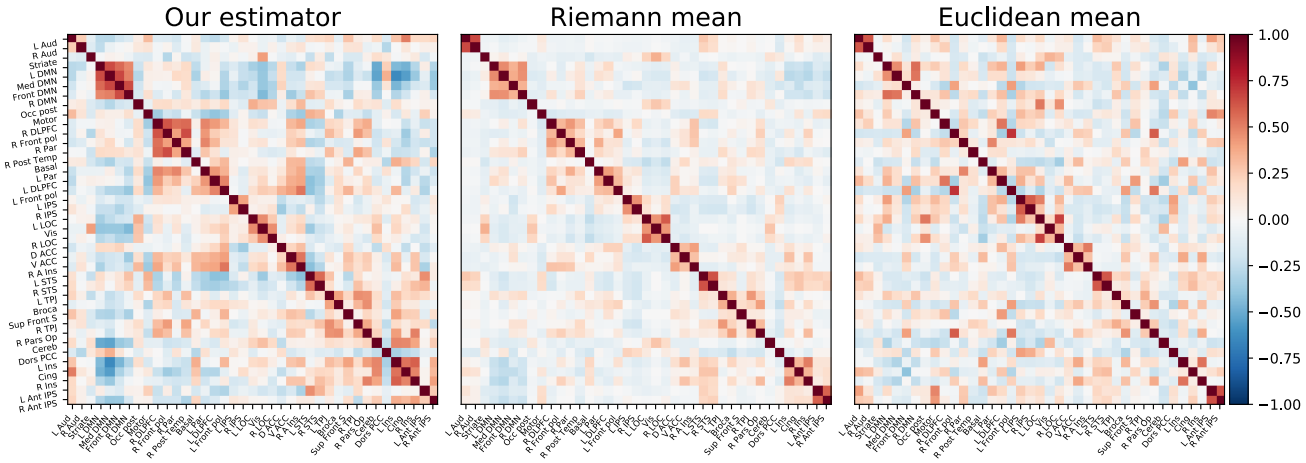

Figure 6. **Comparison of correlation matrices estimated from resting-state fMRI scans of 458 subjects.** Our estimator detects more prominent correlation patterns than the Riemannian and Euclidean mean estimators, especially for the known brain networks (organized in clusters around the main diagonal, see zoomed versions in Fig. 7).

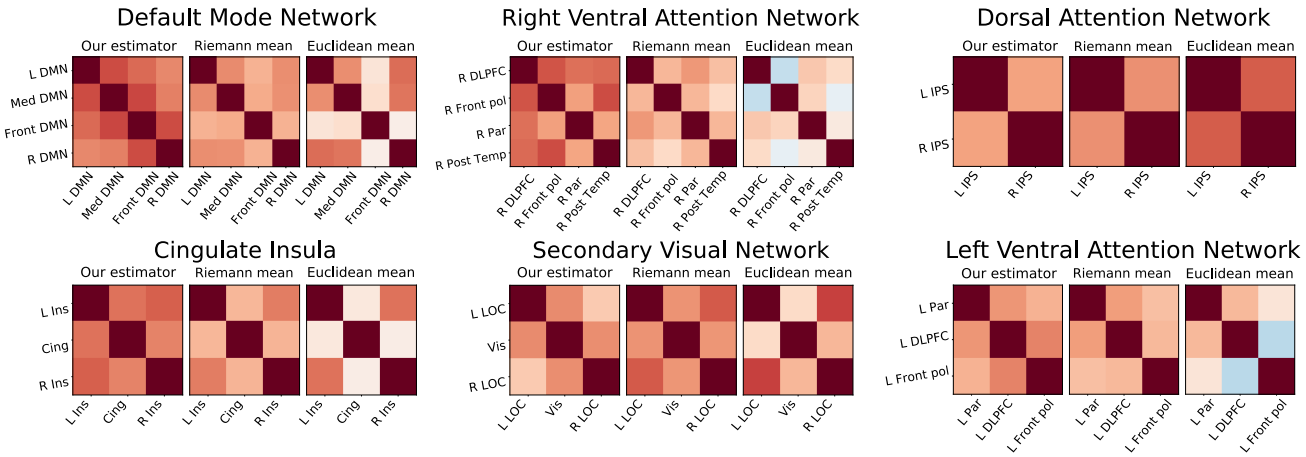

Figure 7. **Zoom in on several known brain networks.** Our estimator detects higher correlations within most known brain networks than the Riemannian mean and the Euclidean mean estimators.

$[0, 2\pi]$, respectively (Chopin, 2011).

Figure 5 depicts the estimation results obtained with Alg. 1 and with naive averaging, using 1000 subjects. We show results for three different common covariance matrices. These include a zero matrix (first row), an identity matrix (second row), and a random PSD matrix (third row). As can be seen, the Euclidean and Riemannian means produce inaccurate estimates in all entries of the matrix while our estimator produces accurate results. This is despite the preference of the noise covariances towards specific patterns.

5.3. FMRI data

Next, we applied our covariance estimation algorithm on the ADHD200-preprocessed dataset (Bellec et al., 2017). We used the Athena pipeline. In particular, we used preprocessed resting state fMRI data, written into MNI space at $4\text{mm} \times 4\text{mm} \times 4\text{mm}$ voxel resolution. We removed nuisance variance (Lund, 2001; Fox et al., 2005), applied a temporal bandpass filter ($0.009\ \text{Hz} < f < 0.08\ \text{Hz}$) (Fox et al., 2005; Biswal et al., 1995; Cordes et al., 2001) and a spatial Gaussian filter (6mm FWHM), and removed linear trend from the extracted time-courses. We took the 458 control subjects from the published training set (for results on 141 subjects with ADHD, please see the Supplementary). From each subject, we extracted time-courses of 39 regions of interest (ROI) of the MSDL atlas (Varoquaux et al., 2011) and estimated their covariance using the Ledoit-Wolf estimator (Ledoit & Wolf, 2004). This gave us a 39×39 covariance matrix per subject. We estimated the common covariance matrix using Alg. 1, using Geometric (Riemannian) mean (Varoquaux et al., 2010a), and using Euclidean mean. From the estimated covariances, we calculated correlation matrices. We used the nilearn and scikit-learn python packages

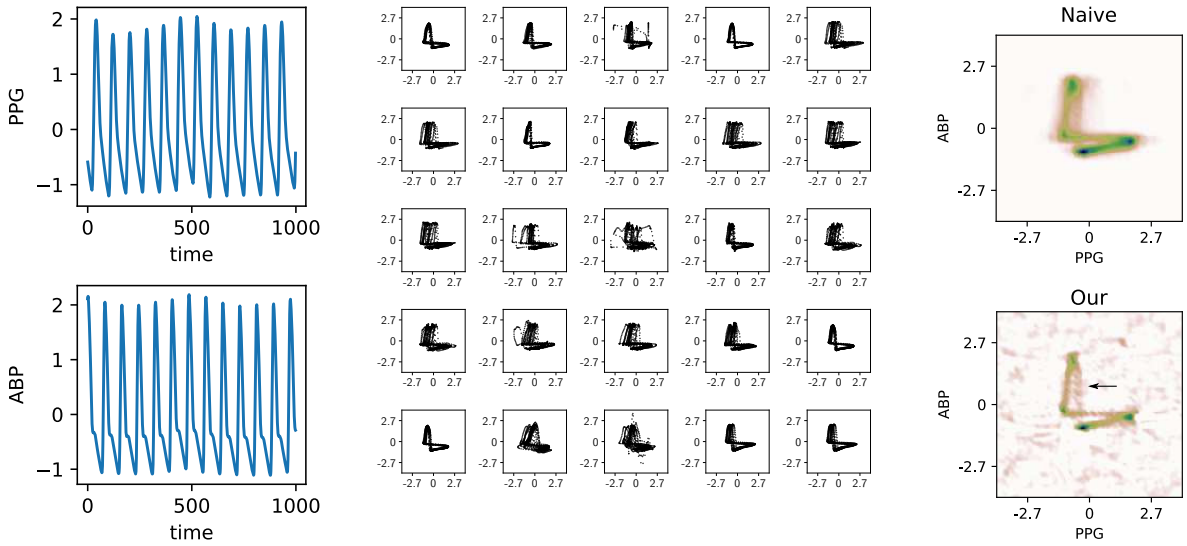

Figure 8. **Estimation of the common joint pdf of PPG and ABP from a group of 25 subjects.** PPG and ABP signals exhibit synchronized variations, as seen on the left for a single subject. Thus, scatter plots of ABP vs. PPG show a unique pattern. However, as can be seen on the middle, the pattern of each subject deviates from the common structure in a different way. Our common pdf estimator manages to unveil delicate common structures, despite those variations. Naive application of KDE on the samples from all subjects, fails to detect those structures.

(Abraham et al., 2014; Pedregosa et al., 2011; Buitinck et al., 2013). The running time of Alg. 1 was about 10s on an 8 core Intel i7-6700 with 16GB of RAM working at 3.40GHz. The results are depicted in Fig. 6.

It has been shown that estimates of connectivity patterns often vary significantly between subjects (Moussa et al., 2012). As can be seen in Fig. 6, our estimator detects activity within known networks despite the large variability between subjects. In particular, our estimator detects stronger correlations than the Euclidean and Riemannian mean estimators within the Default Mode Network, the Right Ventral Attention network, the Left Ventral Attention network, and the Cingulate Insula (connectivity between cingulate cortex and insula) (Moussa et al., 2012). Zoomed versions of those networks are shown in Fig. 7. Note that the Euclidean mean estimator shows very low correlations within some of those regions.

5.4. Common density of PPG and ABP

In our last experiment, we used Alg. 2 to estimate the joint density function of arterial blood pressure (ABP) and photoplethysmogram (PPG) recordings. We used measurements from 25 subjects in critical care taken from the MIMIC 2 dataset (Kachuee et al., 2015). As a preprocessing step, we normalized the signals to have zero mean and unit variance. For each subject, we then estimated the 2D pdf of ABP and PPG using Gaussian KDE with bandwidth 0.08. From the resulting 25 pdfs, we estimated the common pdf using Alg. 2. As can be seen in Fig. 8, our algorithm manages to reveal delicate structures in the common pdf, which are not seen when applying KDE on all the data from all the subjects. In the Supplementary, we show that these structures are not detected with naive KDE *with any bandwidth*. This illustrates again the ability of our approach to suppress subject-specific noise factors that have different distributions.

6. Conclusion

We presented algorithms for estimating the covariance and the pdf of the common component of a group of subjects, when noise has a different distribution for each subject. Our algorithms take advantage of the diversity of the subject-specific noise distributions in order to efficiently suppress them. In contrast to previous approaches, we did not assume any parametric model for the underlying distributions. We proved that under rather mild assumptions, our common covariance estimate tends to the covariance of the common component as the number of subjects grows. We presented experiments on simulated and on real data, which confirmed the advantages of our methods over alternative approaches.

References

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., and Varoquaux, G. Machine learning for neuroimaging with scikit-learn. *Frontiers in neuroinformatics*, 8:14, 2014.

Beckmann, C. F. and Smith, S. M. Tensorial extensions of independent component analysis for multisubject fmri analysis. *Neuroimage*, 25(1):294–311, 2005.

Bellec, P., Chu, C., Chouinard-Decorte, F., Benhajali, Y., Margulies, D. S., and Craddock, R. C. The neuro bureau adhd-200 preprocessed repository. *Neuroimage*, 144:275–286, 2017.

Biswal, B., Zerrin Yetkin, F., Haughton, V. M., and Hyde, J. S. Functional connectivity in the motor cortex of resting human brain using echo-planar mri. *Magnetic resonance in medicine*, 34(4):537–541, 1995.

Blau, Y. and Michaeli, T. Non-redundant spectral dimensionality reduction. In *ECML/PKDD*, 2017.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., and Varoquaux, G. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pp. 108–122, 2013.

Bullmore, E., Brammer, M., Williams, S. C., Rabe-Hesketh, S., Janot, N., David, A., Mellers, J., Howard, R., and Sham, P. Statistical methods of estimation and inference for functional mr image analysis. *Magnetic Resonance in Medicine*, 35(2):261–277, 1996.

Calhoun, V. D., Adali, T., Pearlson, G. D., and Pekar, J. A method for making group inferences from functional mri data using independent component analysis. *Human brain mapping*, 14(3):140–151, 2001.

Chen, G., Saad, Z. S., Britton, J. C., Pine, D. S., and Cox, R. W. Linear mixed-effects modeling approach to fmri group analysis. *Neuroimage*, 73:176–190, 2013.

Chopin, N. Fast simulation of truncated gaussian distributions. *Statistics and Computing*, 21(2):275–288, 2011.

Cordes, D., Haughton, V. M., Arfanakis, K., Carew, J. D., Turski, P. A., Moritz, C. H., Quigley, M. A., and Meyerand, M. E. Frequencies contributing to functional connectivity in the cerebral cortex in resting-state data. *American Journal of Neuroradiology*, 22(7):1326–1333, 2001.

Delbracio, M. and Sapiro, G. Removing camera shake via weighted fourier burst accumulation. *IEEE Transactions on Image Processing*, 24(11):3293–3307, 2015.

Fletcher, P. T. and Joshi, S. Riemannian geometry for the statistical analysis of diffusion tensor data. *Signal Processing*, 87(2):250–262, 2007.

Förstner, W. and Moonen, B. A metric for covariance matrices. In *Geodesy-The Challenge of the 3rd Millennium*, pp. 299–309. Springer, 2003.

Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Van Essen, D. C., and Raichle, M. E. The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(27): 9673–9678, 2005.

Friedman, J., Hastie, T., and Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Friston, K. J., Stephan, K. E., Lund, T. E., Morcom, A., and Kiebel, S. Mixed-effects and fmri studies. *Neuroimage*, 24(1):244–252, 2005.

Hagler, D. J., Saygin, A. P., and Sereno, M. I. Smoothing and cluster thresholding for cortical surface-based group analysis of fmri data. *Neuroimage*, 33(4):1093–1103, 2006.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., and Pietrini, P. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430, 2001.

Kachuee, M., Kiani, M. M., Mohammadzade, H., and Shabany, M. Cuff-less high-accuracy calibration-free blood pressure estimation using pulse transit time. In *Circuits and Systems (ISCAS), 2015 IEEE International Symposium on*, pp. 1006–1009. IEEE, 2015.

Ledoit, O. and Wolf, M. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

Lund, T. E. fcmrimapping functional connectivity or correlating cardiac-induced noise? *Magnetic resonance in medicine*, 46(3):628–628, 2001.

Marrelec, G., Krainik, A., Duffau, H., Pélégrini-Issac, M., Lehéricy, S., Doyon, J., and Benali, H. Partial correlation for functional brain interactivity investigation in functional mri. *Neuroimage*, 32(1):228–237, 2006.

Mensch, A., Varoquaux, G., and Thirion, B. Compressed online dictionary learning for fast resting-state fmri decomposition. In *Biomedical Imaging (ISBI), 2016 IEEE*

13th International Symposium on, pp. 1282–1285. IEEE, 2016.

Moussa, M. N., Steen, M. R., Laurienti, P. J., and Hayasaka, S. Consistency of network modules in resting-state fmri connectome data. *PloS one*, 7(8):e44428, 2012.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Power, J. D., Cohen, A. L., Nelson, S. M., Wig, G. S., Barnes, K. A., Church, J. A., Vogel, A. C., Laumann, T. O., Miezin, F. M., Schlaggar, B. L., et al. Functional network organization of the human brain. *Neuron*, 72(4): 665–678, 2011.

Ramsey, J. D., Hanson, S. J., Hanson, C., Halchenko, Y. O., Poldrack, R. A., and Glymour, C. Six problems for causal inference from fmri. *neuroimage*, 49(2):1545–1558, 2010.

Smith, S. M., Miller, K. L., Salimi-Khorshidi, G., Webster, M., Beckmann, C. F., Nichols, T. E., Ramsey, J. D., and Woolrich, M. W. Network modelling methods for fmri. *NeuroImage*, 54(2):875–891, 2011.

Varoquaux, G. and Craddock, R. C. Learning and comparing functional connectomes across subjects. *NeuroImage*, 80: 405–415, 2013.

Varoquaux, G., Baronnet, F., Kleinschmidt, A., Fillard, P., and Thirion, B. Detection of brain functional-connectivity difference in post-stroke patients using group-level covariance modeling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 200–208. Springer, 2010a.

Varoquaux, G., Gramfort, A., Poline, J.-B., and Thirion, B. Brain covariance selection: better individual functional connectivity models using population prior. In *Advances in neural information processing systems*, pp. 2334–2342, 2010b.

Varoquaux, G., Sadaghiani, S., Pinel, P., Kleinschmidt, A., Poline, J.-B., and Thirion, B. A group model for stable multi-subject ica on fmri datasets. *Neuroimage*, 51(1): 288–299, 2010c.

Varoquaux, G., Gramfort, A., Pedregosa, F., Michel, V., and Thirion, B. Multi-subject dictionary learning to segment an atlas of brain spontaneous activity. In *Biennial International Conference on Information Processing in Medical Imaging*, pp. 562–573. Springer, 2011.